IDIOMATIC SENTENCE GENERATION AND PARAPHRASING

BY

JIANING ZHOU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Adviser:

Assistant Professor Suma Bhat

ABSTRACT

Idiomatic expressions (IE) play an important role in natural language, and have long been a "pain in the neck" for NLP systems. Despite this, text generation tasks related to IEs remain largely under-explored. In this study, we propose two new tasks of idiomatic sentence generation and paraphrasing to fill this research gap. We introduce a curated dataset of 823 IEs, and a parallel corpus with sentences containing them and the same sentences where the IEs were replaced by their literal paraphrases as the primary resource for our tasks. We benchmark existing deep learning models, which have state-of-the-art performance on related tasks using automated and manual evaluation with our dataset to inspire further research on our proposed tasks. By establishing baseline models, we pave the way for more comprehensive and accurate modeling of IEs, both for generation and paraphrasing.

Inspired by psycholinguistic theories of idiom use in one's native language, we also propose a novel approach for these tasks, which retrieves the appropriate idiom for a given literal sentence, extracts the span of the sentence to be replaced by the idiom, and generates the idiomatic sentence by using a large pretrained language model to combine the retrieved idiom and the remainder of the sentence. For idiomatic sentence paraphrasing, the definition of the idiom in the given idiomatic sentence is first retrieved. Then the idiom in the sentence is extracted and finally the literal counterpart is generated by a large pretrained language model. Experiments on a novel dataset created for these tasks show that our model is able to work effectively. Furthermore, automatic and human evaluations show that for these tasks, the proposed model outperforms a series of competitive baseline models for text generation.

Being able to generate literal counterparts of high quality, our method for idiomatic sentence paraphrase is also used for constructing a larger corpus with the help of MAGPIE dataset. This enlarged corpus also helps to improve the performance of different models on idiomatic sentence generation.

ACKNOWLEDGMENTS

This Master journey is truly precious for me. I am really happy that after these two years, I have become a more persistent person and more passionate about research.

First and foremost, I am grateful to my advisor Assistant Professor Suma Bhat. Professor Bhat kindly accepted my as a member in her research group when I was naive about research and had a weak background. I have learned a lot about research from her. Besides, I also appreciate that she was very patient with me. Thanks Suma for all the support.

Next, I would like to thank my colleague, Hongyu Gong, in the group. She really helped a lot with my research. I have learned a lot from her during the discussion with her. I am really grateful to her for being my colleague and friend.

Finally, I would like to thank my parents for their unconditional love and support. Words cannot express my appreciation for what they have done for me since my birth. I know that they are always proud of me no matter what I achieve.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Idiomatic expressions (IEs) make language natural. These expressions, more broadly called a multiword expressions (MWEs) are (non-compositional) phrases whose meaning differs from the literal meaning of their constituent words taken together [1]. Their use imparts naturalness and fluency [2, 3, 4, 5], is prompted by pragmatic and topical functions in discourse [6] and often conveys a nuance in expression (stylistic enhancement) using imagery that is beyond what is available in the context [1]. Idiomatic expressions, including phrasal verbs (e.g., carry out), idioms (e.g., pull one's leg) are also an essential part of a native speakers vocabulary and lexicon [7].

IEs constitute a ubiquitous part of daily language and social communication, primarily used in conversation, fiction and news [8], frequently used by teachers when presenting their lessons to students [9] and occur cross-lingually [1, 10]. Their non-compositionality is the reason for their classical standing as "a pain in the neck" [11] and "hard going" [12] for NLP.

The Oxford English dictionary defines the phrasal verb (an IE) *vote out* as 'To turn (a person) out of office.' Using Google translate[1] to translate the topical slogan "vote them out!" into eight of the world's most spoken and relatively resource-rich languages yielded the results shown in Figure 1.1. As native speakers will attest, other than in Spanish, all the translations mean just the opposite, "vote for them!" This, and other studies on computational processing of idioms and metaphors in [13, 14, 15] reinforce the need for nuanced language processing—a grand challenge for NLP systems.

Gaining a deeper understanding of IEs and their literal counterparts is an important step toward this goal. In this study, we introduce two novel tasks related to paraphrasing between literal and idiomatic expressions in unrestricted text: (1) Idiomatic sentence simplification (ISS) to automatically paraphrase idiomatic expressions in text, and 2) Idiomatic sentence generation (ISG) to replace a literal phrase in a sentence with a synonymous but more vivid phrase (e.g., an idiom). ISS directly addresses the need for performing text simplification in several application settings, including summarizers [16] and parsing [17]. Moreover, ISS may actually be helpful when an idiomatic expression does not have an exact counterpart in a target language. This is akin to the 'translation by paraphrase' strategy recommended for human translation when the source language idiom is obscure and non-existent in the target language [18]. On the other hand, ISG advances the area of text style transfer [19, 20] bringing the as yet unexplored dimension of nuanced language to style transfer.

A second important component of this study is the introduction of a new curated dataset

[1] `https://translate.google.com/`. Accessed November 19, 2020

English	Vote them out!
Spanish	¡Vote para sacarlos!
Arabic	!التصويت لهم
Chinese	投票给他们!
Hindi	उन्हें वोट दें!
French	Votez-les!
German	Stimmen Sie sie ab!
Korean	투표하세요!
Russian	Проголосуйте за них!

Figure 1.1: State-of-the-art machine translations of "Vote them out!" into different languages mean the opposite.

of parallel idiomatic and literal sentences, created for the purpose of advancing progress in nuanced language processing and serving as a testbed for the proposed tasks. Recent literature has explored several aspects of figurative and nonliteral language processing, including detecting and interpreting metaphors [15, 21], disambiguating IEs for their figurative or literal in a given context [17, 22, 23] and analyzing sarcasm [24, 25, 26], by using curated datasets of sentences with linguistic processes in the wild. These datasets are ill-suited for the proposed tasks because they consist of specific figurative constructions (metaphors) [27], do not cover multiple IEs [28, 29], or are not parallel [22, 30] underscoring the need for a new dataset.

The newly constructed dataset permits us to benchmark the performance of several state-of-the-art neural network architectures (seq2seq and pretrained+fine-tuned models, with and without copy-enrichment) that have demonstrated competitive performance in the related tasks of simplification, and style transfer. Using automatic and manual evaluations of the outputs for the two tasks, we find that the existing models are inadequate for the proposed tasks. The sequence-to-sequence models clearly suffer from data sparsity, the added copy mechanism helps preserve the context that is not replaced, and despite their prior knowledge of the pretrained models, they are still limited in their ability to paraphrase and generate. This leads us to discussing novel insights, applications and future directions for related research.

The non-literal and stylized meaning of multi-word expressions (MWE) in general and idioms in particular, pose two broad kinds of challenges. First, they affect readability in target populations. For instance, despite their intact structural language competence, individuals with Asperger syndrome and more broadly those with autism spectrum disorder are known to experience significant challenges understanding figurative language (idioms)

in their native language [31]. It is also widely acknowledged that idiomatic expressions are some of the hardest aspects of language acquisition and processing for second language learners [32, 33, 34]. Second, natural language processing systems are known to be negatively impacted by idioms that are naturally part of text. This has been demonstrated in [13, 14, 15] that highlighted how idioms and metaphors affect machine translation leading to awkward or incorrect translations from English to other languages.

As native speakers can attest, the translated expressions with literal translations of the words of the idioms demonstrate the lack of ability of today's machine translation to handle idioms that are commonplace in language. Other studies have documented the need for text simplification of idioms in several applications including summarizers [16], parsing [17].One way of mitigating these challenges is to create idiom-aware applications that automatically paraphrase idioms in text as literal expressions and enable its subsequent processing by humans[2] and NLP systems, broadly aligned with the human-oriented and the machine-oriented controlled language rule sets proposed by O'Brien [37]. This motivates the first task of replacing an idiom with its literal paraphrase or paraphrasing idiomatic expressions.

Text style transfer is a recent task that has received much attention with regard to sentiment manipulation and writing styles [19, 20]. Automatically replacing a common phrase with a related but more vivid phrase (e.g., an idiom) to serve as a rhetorical device that draws attention through its novel phrasing and lively imagery naturally extends this research direction. This is the second task that we propose—to automatically transform a literal expression into an idiomatic one.

Recent literature has explored several aspects of figurative and nonliteral language processing. These have primarily concerned with detecting and interpreting metaphors [15, 21], disambiguating figurative meaning from literal meaning in context [17, 22, 23] and identifying and analyzing sarcasm [24, 25, 26]. Given the recent advances in paraphrasing, text simplification, text generation and style transfer, our goal is to draw the community's attention to the related but largely under-explored problems of processing idiomatic expressions. In order to spur research in the proposed tasks, this study summarizes our efforts of creating a large dataset of 823 commonly encountered idioms, their definitions and sentences where the idioms have been used, while also including corresponding sentences with the idioms replaced by literal phrases.

The main contributions of this work are summarized as follows.

[2]The Plain Language Action and Information Network (PLAIN) and the European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability, Make it Simple [35] call for making access to information more equitable by making text more accessible to people with learning disabilities and those with reduced language competence [36].

1. We propose two new tasks related to idiomatic expressions—idiomatic sentence simplification and idiomatic sentence generation;

2. We introduce a curated dataset of 823 idiomatic expressions, replete with sentences containing these IEs in the wild and the same sentences where the IEs were replaced by their literal paraphrases.

3. We use the combination of the new dataset and the proposed tasks as a lens through which we gain novel insights about the capabilities of deep learning models for processing nuanced language generation and paraphrasing.

4. We use our method on the MAGPIE dataset to construct a larger dataset with parallel idiomatic sentences and corresponding literal sentences. This enlarged corpus is then used to improve the performance of different models on idiomatic sentence generation.

CHAPTER 2: TASK DEFINITION

We propose two new tasks: **idiomatic sentence generation** transforms a literal sentence into a sentence involving idioms. Used frequently in everyday language, idioms are known to add color to expressions and improve the fluency of communication. The idiomatic rewriting improves the quality of text generation in that it could enhance the textual diversity and convey abstract and complicated ideas in a succinct manner. For example, the idiomatic sentence *BP cut corners and violated safety requirements.* conveys the same idea as its literal counterpart *BP saved time, money and energy and violated safety requirements*, but in a more vivid and succinct manner.

The second task is **idiomatic sentence paraphrasing**, simplifying sentences with idioms into literal expressions. As an example, the sentence–*It is certainly not a sensible move to cut corners with national security*– has the idiom *cut corners* replaced the literal counterpart *save money*. By paraphrasing the idioms from which machine translation often suffers, our task of idiomatic sentence paraphrasing can also benefit machine translation.

In this work, we distinguish our task of idiomatic sentence generation from idiom generation. While the latter task creates new idioms with novel word combinations, our study is to use existing idioms in a sentence and preserve the semantic meaning.

The task of idiomatic sentence paraphrasing is closely related to text simplification that has mostly been studied as related tasks of lexical paraphrasing and syntactic paraphrasing [38]. A significant departure of this task from that of these related tasks that centrally address style is that (i) we aim for local synonymous paraphrasing by transforming not the entire sentence but a phrase in the sentence, (ii) the transformation is not related to syntactic structures, but related to the complexity in meaning[1]. We propose doing joint monolingual translation with simplification and is similar in spirit to [39].

There are many technical challenges to performing these tasks. The task of idiomatic sentence paraphrasing involves first identifying that an expression is an idiom and not a literal expression (e.g. *black sheep*) [23, 29, 40]. Once identified, the IE may have multiple senses (e.g. *tick off*) and its appropriate sense will need to be identified before paraphrasing it. Third, an appropriate literal phrase will have to be generated to replace the IE. Finally, the literal phrase will have to be fit in the surrounding sentential context for a fluent construction. For idiomatic sentence generation, the context of the literal phrase could permit more than one candidate idiom (e.g. keep quiet). In this study, we assume that we have an idiomatic sentence and leave it to future work to explore the task in conjunction with this step.

[1]The consideration of whether idioms are semantic- or pragmatic- or discourse-level phenomena is important, but beyond the scope of this study.

CHAPTER 3: RELATED WORK

The theme of this study is naturally connected to three streams of text generation tasks—paraphrasing, style transfer and metaphoric expression generation. We will discuss these tasks and also the datasets used in these tasks to study their similarities and differences to our dataset and tasks.

3.1 PARAPHRASE

The aim of paraphrasing is to rewrite a given sentence while preserving its original meaning. Being widely studied in the recent research, many datasets have been constructed to facilitate the task. PPDB [41], MRPC [1], Twitter URL Corpus [42], Quora [2] and ParaNMT-50M [43] have been the most commonly used datasets. The most commonly used Seq2Seq models have been successfully applied to paraphrasing [44, 45, 46, 47]. Besides the end-to-end models, a template-based pipeline model was proposed to divide paraphrase generation into template extraction, template transforming and template filling [48]. However, unlike paraphrasing a sentence or a literal-to-literal paraphrasing task, our proposed tasks are more constrained given the existence of idiomatic expressions. This renders the datasets used for the task of paraphrasing and the associated paraphrasing models inadequate for our task. Our dataset is created to fill this need to advance a fundamental understanding of idiomatic text generation and paraphrasing. Therefore, research into our tasks and dataset can also be used for paraphrasing when only part of the sentence need to be paraphrased.

3.2 STYLE TRANSFER

The task of style transfer can be defined as rewriting sentences into those with a target style. Recent research has primarily focused sentiment manipulation and changes in writing styles [19, 20]. Our proposed tasks are different from the nature of style transfer studies in recent works because (i) our tasks retain a large portion of the input sentences while style transfer may need to completely change the input sentences, and (ii) our tasks explore the nuance component of style, an aspect heretofore unexplored. To test different models' performance on style transfer, several non-parallel corpora have been used (Yelp [49], Grammarly's Yahoo Answers Formality Corpus [50], Amazon Food Review dataset [51] and Product Review

[1] https://www.microsoft.com/en-us/download/details.aspx?id=52398
[2] https://www.kaggle.com/aymenmouelhi/quora-duplicate-questions

Dataset	Task	Size	# idioms	Sent Len (original)	Sent Len (target)
PIE (ours)	Ours	3,524/823/823	823	18.5	19.0
Para-NMT	Paraphrase	5,370,128	-	11.43	10.56
WikiLarge	Text Simplification	296,402/992/359	-	24.1	15.51
Metaphor	Metaphor Generation	171	-	7.30	7.37

Table 3.1: Comparison of our dataset with other related datasets. Training, validation and testing size splits are provided when applicable. Data in all these datasets is a combination of collection from the wild and manual generation. In our corpus, original sentences are idiomatic sentences and target sentences are literal sentences.

dataset [52]). Despite their size, they lack the focus on IEs and are all non-parallel. This has led to the the study of unsupervised methods for style transfer, including cross-aligned auto-encoder [53], VAE [53], Generative Adversarial Network [54], reinforcement learning for constraints in style transfer [20, 55] and pipeline models [56, 57]. Owing to the essential departure of our tasks from those of previously studied style transfer tasks, and the limitation of non-parallel corpus, we create our own parallel dataset which focuses on IEs.

3.3 METAPHORIC EXPRESSION GENERATION

Prior work on automated metaphor processing has primarily focused on their identification, interpretation and also generation. [15, 21, 58]. Also, data for this task is extremely sparse: there are not any large scale parallel corpora containing literal and metaphoric paraphrases which aims for metaphor generation. The most useful one is that of [59]. However, their dataset has a small number (171) of metaphoric sentences extracted from WordNet. Early works on metaphor generation mainly focus on phrase level metaphor and template-based generation [60, 61]. Recent works also explore the power of neural networks [62, 63, 64]. However, most of the research on metaphor generation suffer from the lack of parallel corpora.

Our proposed tasks share some similarities with metaphor generation but also have differences. Instead of focusing on paraphrase of single word like most metaphor generation work, our tasks often require a mapping between two multi-word expressions, which makes our tasks more challenging.

3.4 TEXT SIMPLIFICATION

Text simplification aims to rewrite input sentences into lexically and/or syntactically simplified forms. The Simple Wikipedia Corpus [65] and more recently, the Newsela dataset

[38] and the WikiLarge dataset [66] dominate the research area. The use of different machine learning models have also been explored for this task, including statistical machine translation model [67], the Seq2Seq architecture [68] and the Transformer architecture [69].

Departing from previous attempts at lexical or syntactic simplification, our proposed task of idiomatic sentence paraphrasing aims to simplify the nuance of non-compositional and figurative expressions thereby permitting a more literal understanding of the sentence.

We summarize the datasets of the related tasks in Table 3.1.

CHAPTER 4: DATASET CONSTRUCTION

We describe the details of the data collection, data annotation, corpus analyses and comparisons with other existing corpora.

4.1 DATA COLLECTION

The Parallel Idiomatic Expression Corpus (PIE), consists of idiomatic expressions (IEs), their definitions, sentences containing the IEs and corresponding sentences where the IEs are replaced with their literal paraphrases. One instance of the dataset is shown in Figure 4.1.

We collected a list of 1042 popular IEs and their meanings from a public educational website [1] that has a broad coverage of frequently used IEs including phrasal verbs, idioms and proverbs. For a broad coverage of IEs we did not limit them to a specific syntactic category. Some IEs such as "tick off" (Figure 4.1) have multiple senses. We labeled the sense of IEs in given sentences according to the sense information from reliable sources including the Oxford English Dictionary[2], the Webster Dictionary [3] and the Longman Dictionary of Contemporary English[4]. IEs that were not available in any of the popular dictionaries were excluded from dataset as were proverbs that are independent clauses (e.g., *the pen is mightier than the sword*). To guarantee each sense is well represented, we collected at least 5 sentences for each sense of an IE from online source.

The data collection step yielded the corpus with a total of 823 IEs and 5170 sentence-pairs using these IEs (an average of 6.3 sentence-pairs per idiom). We also note that every instance (idiomatic-literal pair) is only one sentence long. The corpus statistics are summarized in Table 8.1.

4.2 DATA ANNOTATION

In order to create the parallel dataset of idiomatic and literal sentences for the proposed tasks, we rewrite each idiomatic sentence into its literal form, where the IE was replaced by a literal phrase. As part of this manual paraphrasing, we paraphrase only the IE so as not to alter its meaning in the context of the sentence, preserving the phrases syntactic function and to conform to the sense definition. The rest of the sentence was to be left

[1] www.theidioms.com
[2] https://www.oxfordlearnersdictionaries.com
[3] https://www.merriam-webster.com
[4] https://www.ldoceonline.com

Idiom	Tick Off	
Sense	to complete an item on a list	to make someone angry or offended
Idiomatic Sentence	I would like to tick off some more items on my list before going home	My decision is going to tick off my entire family.
Idiomatic Labels	O O O O **B I** O O O O O O O O O	O O O O O **B I** O O O.
Literal Sentence	I would like to cross out some more items on my list before going home	My decision is going to anger my entire family.
Literal Labels	O O O O **B I** O O O O O O O O O	O O O O O **B** O O O.

Figure 4.1: An example from our dataset. Idioms are highlighted in blue, and their literal paraphrases are in red.

Statistics	**# of instances**	**Avg. # of words**
Idioms	823	3.2
Sense	862	7.9
Idiomatic sent	5170	19.0
Literal sent	5170	18.5

Table 4.1: Statistics of our parallel corpus.

unchanged. Original sense definition can be freely used when rewriting or use paraphrases of sense definition. After the first annotation pass, we checked the literal sentences generated and corrected any errors.

To specify the span of the IE in each idiomatic sentence and that of the literal paraphrase in the corresponding literal sentence, **BIO** labels were used; **B** marks the beginning of the idiom expressions (resp. the literal paraphrases), **I** the other words in the IE (resp. words in the literal paraphrases) and **O** all the other words in the sentences. This labeling was done automatically considering that the only difference between a given idiomatic sentence and its literal sentence is the replacement of idiom with literal phrase. An example of the **BIO** labeled sentence pair is shown in Figure 4.1.

4.3 CORPUS ANALYSES

We summarize the statistics of our PIE dataset in Table 8.1 and compare it with existing datasets in Table 3.1. We notice that the parallel sentences in our dataset are comparable in terms of sentence length, while simple sentences are much shorter in the text simplification dataset. This suggests that the tasks we propose may not result in significantly shorter sentences compared to their inputs, and this constitutes a core departure from the task of

% n-grams	PIE	Para-NMT	Wiki-Large	Metaphor
uni-grams	13.86	46.34	36.2	16.88
bi-grams	23.60	71.24	52.56	36.59
tri-grams	30.19	82.26	58.75	59.61
4-grams	36.51	86.46	62.79	74.41

Table 4.2: The percentage of n-grams in source sentences which do not appear in the target sentences. In our case, it is the percentage of n-grams in literal sentences which do not appear in the idiomatic sentences.

# senses	# of idioms	# pairs	Avg. # of words
1	788	4788	3.2
2	31	322	2.6
3	4	60	2.0

Table 4.3: Statistics of sense distribution. An idiom has an average of 1.05 senses.

text simplification. Moreover, the sentences in our dataset are longer on an average compared to the sentences in existing datasets (with the exception of text simplification data). This can pose challenges to the text generation model performing the tasks proposed in this study.

We also report the percentage of n-grams in the literal sentences which do not appear in the idiomatic sentences as a measure of the difference between the idiomatic and literal sentences. As shown in Table 4.2, there is smaller variation between the source sentences and the target sentences in our dataset. This is again due to the nature of our task, which calls for a local paraphrasing (rewriting only a part of the sentence).

We note that IEs may be naturally ambiguous due to the existence of both figurative and literal senses, as also pointed out in previous works.A small portion of IEs in our dataset have multiple senses, and one example is “tick off ” in Figure 4.1. Table 4.3 presents the distribution of the senses in the IEs in our dataset, and the average number of senses is 1.05, suggesting that the majority IEs in our dataset are monosemous.

4.4 DATASET QUALITY

Noting that the idiomatic to literal sentences were manually created, the quality of our dataset may be called into question. We point out that in an effort to quickly use sentences of good quality and in line with existing datasets for related tasks with idiomatic expressions [29, 30] we collected idiomatic expressions in the wild. However, as acknowledged by previous dataset creation efforts, not all IEs occur equally frequently, which can result in

a representation bias. In addition, finding true paraphrases of IEs in the wild is hard. In light of these practical data-related concerns, we resorted to a manual paraphrasing of the IEs as a trade-off between naturalness and representation. This idea of using non-natural instances is also influenced by successful recent approaches to training data collection and data augmentation using synthetic methods reported in severely resource-constrained domains such as machine translation [70] and clinical language processing [71].

CHAPTER 5: MODEL

5.1 IDIOMATIC SENTENCE GENERATION

The task of idiomatic sentence generation is to rephrase a given literal sentence into its idiomatic counterpart by using an IE to replace a literal phrase while preserving the original meaning of the sentence. This task can be regarded as paraphrasing only a portion of the original sentence because we retain the remaining portion intact. We use ideas about native speakers accessing a mental lexicon of formulaic expressions, including IEs [7, 72, 73, 74] to propose a generation model built upon a pipeline of three modules that perform idiom retrieval, span extraction and idiomatic sentence generation.

An illustration of the pipeline is shown in Fig. 5.2. The input literal sentence is "The visitors headed for shelter when it started to rain ." The idiom retrieval module, using the available idioms and their definitions, retrieves an idiom that fits in this sentence well, which for this example is "run for cover". This idiom will then be used in our generated text. Following this, the span extraction module decides the span of the literal sentence to be replaced with the selected idiom. The selected span is "headed for shelter", a semantic equivalent of the idiom "run for cover". Lastly, the generation module generates the idiomatic sentence based on the retrieved idiom and the input sentence marked with the selected span. Fig. 5.2 shows the generated sentence "The visitor ran for cover when it started to rain .", where the selected span is replaced with the retrieved idiom. We will next elaborate upon each module.

5.1.1 Idiom Retrieval

We use the lexicon with idioms and their definitions created as part of the dataset described in Chapter 4. The module for idiom retrieval searches an idiom that best fits the given literal sentence. It is built upon a pretrained RoBERTa model [75] and a feed-forward classifier. The RoBERTa model takes as input a sequence of tokens, and generates a contextualized representation for each token as well as the whole sequence. The classifier takes the learned representation and predicts whether an idiom fits in well with the given sentence.

Suppose that we have an input literal sentence s, and an idiom i. The retrieval module prepares a token sequence by concatenating a special token "[CLS]", the input literal sentence, the idiom and its definition. The token "[CLS]" is added to the beginning of the sequence in order to comply with the input format of RoBERTa. This sequence is fed to RoBERTa, and

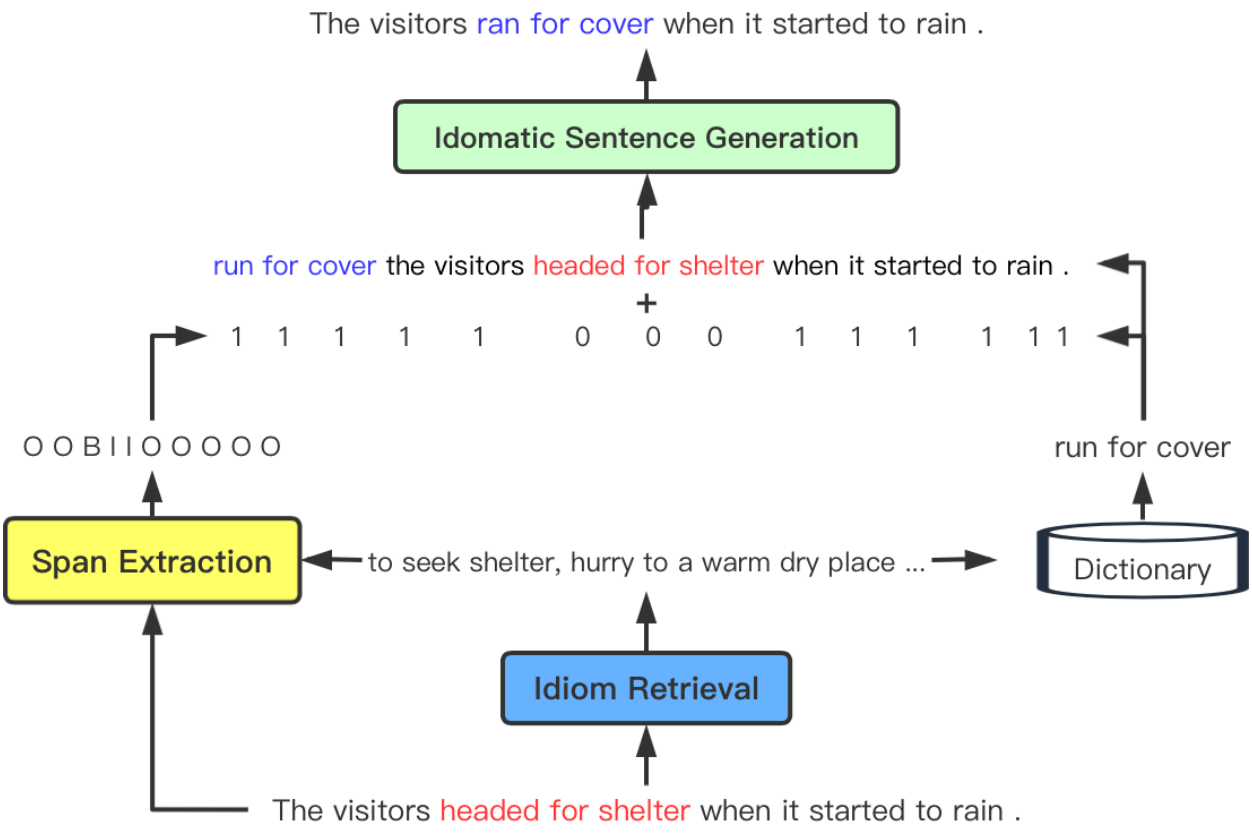

Figure 5.1: The workflow of the pipeline model for idiomatic sentence generation.

we derive the sequence embedding $\mathbf{h}^s_{\text{ret}}(i)$ from the learned representation of each token in the sequence by adding them together.

The feed-forward classifier takes the sequence embedding and outputs a retrieval score $r^s_{\text{ret}}(i)$ to measure how well the idiom i matches sentence s.

$$r^s_{\text{ret}}(i) = \mathbf{W}_{\text{ret}}\mathbf{h}^s_{\text{ret}}(i) + \mathbf{b}_{\text{ret}}, \tag{5.1}$$

where the weight matrix $\mathbf{W}_{\text{ret}}$ and the bias vector $\mathbf{b}_{\text{ret}}$ are parameters of the classifier.

Training. An input instance to the retrieval module was a sentence-idiom pair. An instance was considered as a positive instance and labeled as "1", if the idiom was used to rewrite the literal sentence in the parallel dataset. For each positive instance, we also created negative instances with the same literal sentence by randomly sampling 100 idioms that did not fit with the sentence. These negative instances were labeled as "0". The retrieval module was trained with the cross-entropy loss to classify the label of a sentence-idiom pair. Parameters were tuned for both RoBERTa and the classifier in the retrieval module.

Test. Given a literal sentence s during testing, we created an input instance for every

idiom i in the dictionary. The retrieval module scores each pair (s, i). The idiom i^* with the highest score is then selected for the literal sentence, i.e., $i^* = \underset{i}{\operatorname{argmax}}\, r^s_{\text{ret}}(i)$.

5.1.2 Span Extraction

After selecting the idiom for a given sentence s, we need to decide which phrase of the input literal sentence should be replaced by this idiom. The span extraction module extracts the span of the words of the phrase from the input sentence. We use the ground truth BIO labels marking the span of the phrase in the input sentence (refer to the Dataset section) and cast the span extraction task as a sequence labeling problem.

Our span extractor consists of a RoBERTa model and a classifier based on Conditional Random Field [76]. The RoBERTa model learns the contextualized representations, which are used by the CRF classifier to label each token in the literal sentence with the B, I, O labels.

Since the span to be replaced in the literal sentence is semantically similar to the definition of the idiom , we again prepare the input sequence of the span extraction module by concatenating the literal sentence and definition of the idiom. Suppose that the embedding of token j in sentence s learned by the RoBERTa model is $\mathbf{h}^s_{\text{ext}}(j)$. A CRF classifier jointly predicts the likelihood $\mathbf{p}^s_{\text{ext}}(j)$ over the label set {B, I, O} for each token j in the sentence s. Suppose that sentence s has n tokens.

$$\mathbf{p}^s_{\text{ext}}(1), \ldots, \mathbf{p}^s_{\text{ext}}(n) = \text{CRF}(\mathbf{h}^s_{\text{ext}}(1), \ldots, \mathbf{h}^s_{\text{ext}}(n)), \tag{5.2}$$

where CRF$(\cdot)$ is the CRF-based sequence classifier.

Training. Both RoBERTa and the CRF classifier in the span extractor are trained using a weighted cross-entropy loss. The weighted loss is adopted to mitigate the imbalanced distribution of labels, since the number of label "O" is much larger than that of other labels. The weight is set as 0.48 for the labels "B" and "I" and 0.04 for the others.

Test. The span extractor outputs labels with the highest likelihood for all tokens in the literal sentence. The tokens with the labels "B" or "I" correspond to the span we want to replace.

5.1.3 Idiomatic Sentence Generation

In the generating stage, we combined the results from the retrieval and deletion stages and use a fine-tuned BART model to generate final output— the idiomatic sentences for the task

of idiomatic sentence generation. All the hyper-parameters for the RoBERTa model and the BART model were set to their default values.

5.2 IDIOMATIC SENTENCE PARAPHRASING

Similar to the task of idiomatic sentence generation, the task of idiomatic sentence paraphrasing is to rephrase a given idiomatic sentence into its literal counterpart by using literal phrases to replace the IE while preserving the original meaning of the sentence. This task can be regarded as paraphrasing only a portion of the original sentence because we retain the remaining portion intact. We still use ideas about native speakers accessing a mental lexicon of formulaic expressions, including IEs [7, 72, 73, 74] to propose a generation model built upon a pipeline of three modules that perform idiom retrieval, span extraction and idiomatic sentence paraphrasing.

An illustration of the pipeline is shown in Fig. 5.2. The input idiomatic sentence is "The visitors ran for cover when it started to rain ." The idiom retrieval module, using the available idioms, retrieves the definition, which for this example is "to seek shelter". This definition will then be used in our generated text. Following this, the span extraction module decides the span of the idiom in the idiomatic sentence to be replaced with the selected definition. The selected span is "ran for cover", a semantic equivalent of the phrase "headed for shelter". Lastly, the generation module generates the literal sentence based on the retrieved definition and the input sentence without the selected span. Fig. 5.2 shows the generated sentence "The visitor headed for shelter when it started to rain .", where the selected span is replaced with the literal phrase. We will next elaborate upon each module.

5.2.1 Idiom Retrieval

We use the lexicon with idioms and their definitions created as part of the dataset described in Chapter 4. The module for idiom retrieval searches the definition of the idiom in the idiomatic sentence. It is built upon a pretrained RoBERTa model [75] and a feed-forward classifier. The RoBERTa model takes as input a sequence of tokens, and generates a contextualized representation for each token as well as the whole sequence. The classifier takes the learned representation and predicts whether the definition fits in well with the given sentence.

Suppose that we have an input idiomatic sentence s, and an idiom i. The retrieval module prepares a token sequence by concatenating a special token "[CLS]", the input idiomatic sentence and the idiom. The token "[CLS]" is added to the beginning of the sequence in

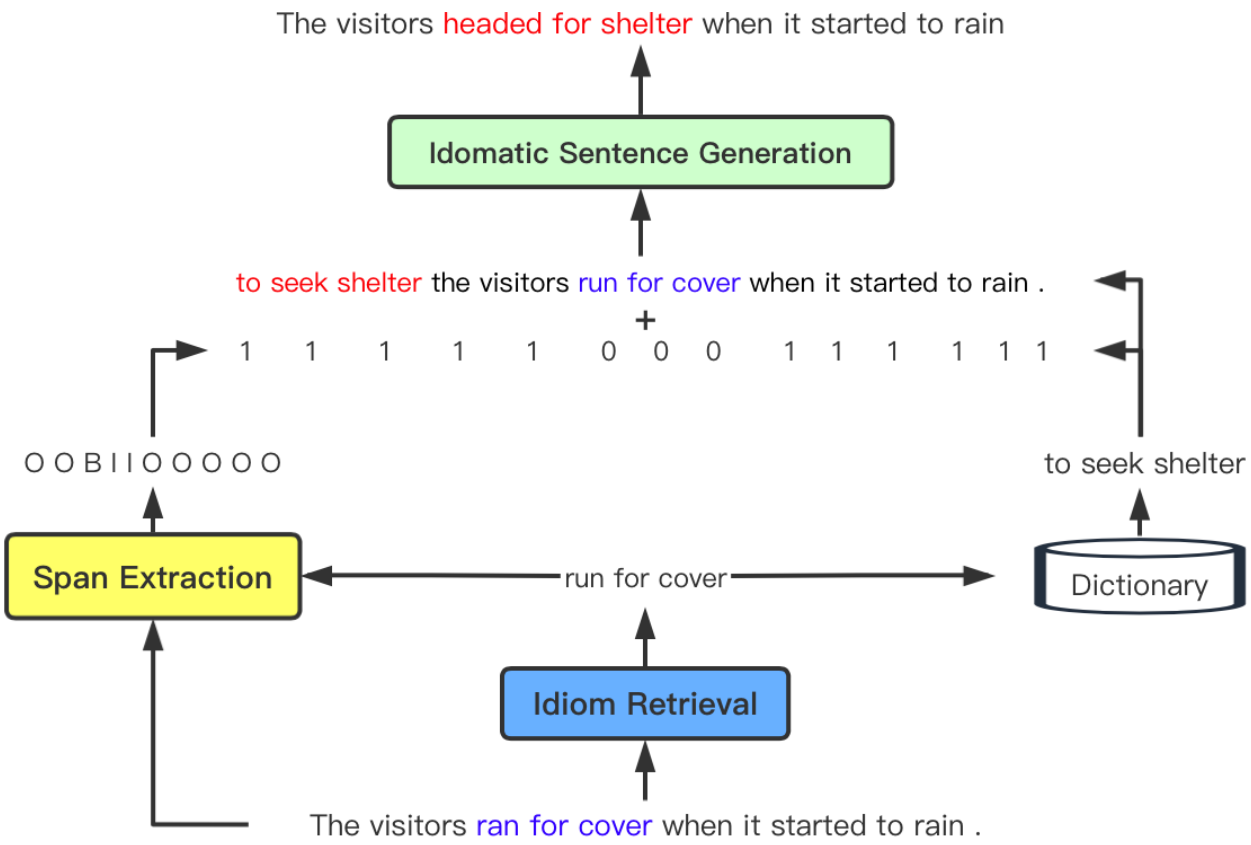

Figure 5.2: The workflow of the pipeline model for idiomatic sentence generation.

order to comply with the input format of RoBERTa. This sequence is fed to RoBERTa, and we derive the sequence embedding $\mathbf{h}^s_{\text{ret}}(i)$ from the learned representation of each token in the sequence by adding them together.

The feed-forward classifier takes the sequence embedding and outputs a retrieval score $r^s_{\text{ret}}(i)$ to measure how well the idiom i matches sentence s.

$$r^s_{\text{ret}}(i) = \mathbf{W}_{\text{ret}}\mathbf{h}^s_{\text{ret}}(i) + \mathbf{b}_{\text{ret}}, \tag{5.3}$$

where the weight matrix $\mathbf{W}_{\text{ret}}$ and the bias vector $\mathbf{b}_{\text{ret}}$ are parameters of the classifier. Finally, the definition is retrieved with the help of a dictionary.

Training. An input instance to the retrieval module was a sentence-idiom pair. An instance was considered as a positive instance and labeled as "1", if the idiom was in the idiomatic sentence in the parallel dataset. For each positive instance, we also created negative instances with the same idiomatic sentence by randomly sampling 100 idioms that are not in the sentence. These negative instances were labeled as "0". The retrieval module was trained with the cross-entropy loss to classify the label of a sentence-idiom pair. Parameters were

tuned for both RoBERTa and the classifier in the retrieval module.

Test. Given an idiomatic sentence s during testing, we created an input instance for every idiom i in the dictionary. The retrieval module scores each pair (s, i). The idiom i^* with the highest score is then selected for the idiomatic sentence, i.e., $i^* = \underset{i}{\operatorname{argmax}}\, r^s_{\text{ret}}(i)$. Finally, the definition is retrieved with the help of a dictionary.

5.2.2 Span Extraction

After selecting the definition for a given sentence s, we need to decide which idiom of the input idiomatic sentence should be replaced. The span extraction module extracts the span of the words of the idiom from the input sentence. We use the ground truth BIO labels marking the span of the idiom in the input sentence (refer to the Dataset section) and cast the span extraction task as a sequence labeling problem.

Our span extractor consists of a RoBERTa model and a classifier based on Conditional Random Field [76]. The RoBERTa model learns the contextualized representations, which are used by the CRF classifier to label each token in the idiomatic sentence with the B, I, O labels.

We prepare the input sequence of the span extraction module by concatenating the idiomatic sentence and the idiom. Suppose that the embedding of token j in sentence s learned by the RoBERTa model is $\mathbf{h}^s_{\text{ext}}(j)$. A CRF classifier jointly predicts the likelihood $\mathbf{p}^s_{\text{ext}}(j)$ over the label set {B, I, O} for each token j in the sentence s. Suppose that sentence s has n tokens.

$$\mathbf{p}^s_{\text{ext}}(1), \ldots, \mathbf{p}^s_{\text{ext}}(n) = \text{CRF}(\mathbf{h}^s_{\text{ext}}(1), \ldots, \mathbf{h}^s_{\text{ext}}(n)), \tag{5.4}$$

where CRF$(\cdot)$ is the CRF-based sequence classifier.

Training. Both RoBERTa and the CRF classifier in the span extractor are trained using a weighted cross-entropy loss. The weighted loss is adopted to mitigate the imbalanced distribution of labels, since the number of label "O" is much larger than that of other labels. The weight is set as 0.48 for the labels "B" and "I" and 0.04 for the others.

Test. The span extractor outputs labels with the highest likelihood for all tokens in the sentence. The tokens with the labels "B" or "I" correspond to the span to be replaced.

5.2.3 Idiomatic Sentence Paraphrasing

In the generating stage, we combined the results from the retrieval and deletion stages and use a fine-tuned BART model to generate final output— the literal sentences for the task of

idiomatic sentence paraphrasing. All the hyper-parameters for the RoBERTa model and the BART model were set to their default values.

CHAPTER 6: EXPERIMENTS

6.1 BASELINES

Considering that our tasks of idiomatic sentence generation and paraphrasing have never been studied before and the fact that they are both text generation tasks, we first choose some basic end-to-end models which have shown state-of-the-art performance on other related text generation tasks. Accordingly, we used the following baselines alluded to as the **models that translate.**:

- **Seq2Seq Model:** an encoder-decoder model built on Long Short Term Memory (LSTM), which is used in neural machine translation [77].
- **Transformer:** a deep neural network with self-attention mechanism [78].

Based on the observation that the idiomatic sentences and literal sentences share much of the context, which remains unchanged during generation, we also use the following **models that copy**:

- **Seq2Seq Model with Copy Mechanism:** an LSTM-based Seq2Seq model which is able to copy directly from inputs [19].
- **Transformer with Copy Mechanism:** a Transformer-based Seq2Seq model which is able to copy directly from inputs [79] [1].

Moreover, considering the similarity between our tasks and paraphrasing, we also choose the **pretrained BART** [80], used for text simplification and paraphrasing, which was fine-tuned on our training instances.

6.2 EXPERIMENTAL SETUP

Baseline Models: For the models that translate and the models that copy, the dimension of the hidden state vectors was set to 256 and the dimension of the word embeddings to 256. The batch size and base learning rates were set to 32 and $1e-3$. These baselines were trained with the parallel sentence pairs as appropriate, i.e., take literal sentences as input and generate the corresponding idiomatic sentences or vice versa. For the pretrained BART model, all the hyper-parameters are set to default.

[1] https://github.com/lipiji/TranSummar

Pipeline Model: Novel instances of idiomatic sentences cannot be generated without previously encountering the IE. Considering this, we set up the pipeline model with the retrieval stage to retrieve an IE for a given literal sentence (resp. the correct idiom given an idiomatic sentence). A RoBERTa model for sentence classification was fine-tuned on our training data. The concatenation of input sentence and correct idiom or sense is labeled by 1 and concatenation of input sentence and irrelevant idioms or senses is labeled by 0. Given all the concatenations of the input sentence and the idioms in our dataset, this stage aims to classify the correct one. In the deletion stage, we deleted the literal phrase that should have been replaced by the retrieved idioms (resp. deleted the IE in the given idiomatic sentence). Again, a RoBERTa model for sequence classification was fine-tuned on our training data with **BIO** labels. This stage aims to assign one of the **BIO** labels for each token in the input sentence and delete the tokens with labels of **B** and **I**. In the generating stage, we combined the results from the retrieval and deletion stages and use a fine-tuned BART model to generate final output— the literal sentences for the task of idiomatic sentence paraphrasing and idiomatic sentences for the task of idiomatic sentence generation. All the hyper-parameters for the RoBERTa model and the BART model were set to default.

6.3 EVALUATION

For automatic evaluation, Rouge [81], BLEU [82], METEOR [83] and SARI [84] are used to compare the similarity between the generated sentences and the references. These metrics has been widely used in various text generation tasks such as paraphrasing, style transfer and text simplification. To measure linguistic quality, we use a pre-trained language model BERT to calculate perplexity scores and a recently proposed measure, GRUEN [85].

Considering that automatic evaluation cannot fully analyze the results, we use human evaluation as a complement to the automatic evaluation metrics. For each task, We randomly sampled 100 input sentences and the corresponding outputs of all baselines. Human annotations were collected with respect to context, style and fluency of generated sentences based on the following criteria.

(1) **Context preservation** measures how well the context surrounding the idiomatic/literal phrase is preserved in the output.

(2) **Target inclusion** checks whether the correct IE or literal phrase is used in the output.

(3) **Fluency** evaluates the fluency and readability of the output sentence including how appropriately the verb tense, noun and pronoun forms are used. (4) **Overall meaning** evaluates the overall quality of the output sentence.

For each output sentence, two annotators with native-speaker-level English proficiency

were asked to rate it on a scale from 1 to 6 in terms of the context preservation, fluency and overall meaning. Higher scores indicate better quality. As for the target inclusion, they were asked to rate it on a scale from 1 to 3. Score 1 denotes that the target phrase is not included in the input at all, 2 denotes partial inclusion, and 3 is for the complete inclusion. We report the average score over all samples for each baseline in each aspect.

CHAPTER 7: RESULTS

Results. We report the automatic and human evaluation results in Table 7.2 and 7.1. More detailed results with all the metrics considered are in the appendix. On both tasks, going by the automatic metrics, copy-enriched transformer, pretrained BART model and the pipeline model perform better than other baselines. Pretrained BART achieved the best performance in BLEU and GRUEN, and the pipeline model does best in SARI. As for human evaluation, BART and the pipeline again achieve the best performance among the baselines. While BART is the best in preserving contexts and achieving fluency, the pipeline is the best in idiom paraphrasing and generation.

Model competence. BART and the pipeline model outperform other baselines in that they leverage auxiliary information (large pretraining corpora and selective idiomatic expression information, respectively) which is not available to the other models. The benefit of the copy mechanism by explicitly retaining the contexts as required by our tasks, is shown in the corresponding gains in automatic and manual evaluation scores for both Seq2Seq and transformer models.

When it comes to the comparison between BART and the pipeline, BART does better in retaining the contexts surrounding idiomatic expressions given its high context score in human evaluation while the pipeline is better at handling the idiomatic part, i.e., target inclusion. Despite the reported superior performance of BART in related text generation tasks [80], our experiments show that BART has limited capability in idiom paraphrasing and generation. The pipeline method, by virtue of error propagation from its retrieval and deletion modules suffers in terms of both the context preservation and fluency. For task of idiomatic sentence generation, the accuracy for retrieval module is 0.27 and F1 score for deletion module is 0.68. For task of idiomatic sentence paraphrasing, the accuracy for retrieval module is 0.96 and F1 score for deletion module is 0.85.

Comparison between two tasks. According to human evaluation results in Table 7.1, both BART and the pipeline received higher scores for idiomatic sentence paraphrasing than idiomatic sentence generation, suggesting that paraphrasing is relatively easier among the two tasks. This resonates with our intuitions as language users in that given a lexical resource, paraphrasing an IE is easier than finding the right IE to replace a phrase.

Limitation of automatic metrics. Table 7.3 presents the correlation between automatic metrics and human judgements. All the correlation scores between automatic metrics and human evaluate scores are not high enough. For BLEU and SARI which mainly measure overlapping tokens, some synonymous idioms or literal phrases are ignored while they are still

Model	Context		Target		Fluency		Overall	
	s2i	i2s	s2i	i2s	s2i	i2s	s2i	i2s
Seq2Seq	1.3	1.2	1.1	1.1	1.1	1.0	1.7	1.7
Seq2Seq with copy	3.8	3.8	1.6	1.7	2.1	3.4	3.5	3.6
Transformer	4.2	4.3	1.3	1.2	3.3	3.4	3.4	3.3
Transformer with copy	5.4	5.3	1.2	1.6	4.6	4.6	3.9	4.2
Pretrained BART	**5.9**	**5.9**	1.5	2.1	**5.9**	**5.9**	4.4	5.0
Pipeline	5.6	5.8	**1.7**	**2.2**	5.1	5.3	**4.5**	**5.1**

Table 7.1: Human evaluation results for the two tasks.

Model	BLEU		SARI		GRUEN	
	s2i	i2s	s2i	i2s	s2i	i2s
Seq2Seq	25.16	42.96	24.13	33.89	32.25	33.45
Seq2Seq with copy	38.02	47.58	43.02	49.69	27.79	32.84
Transformer	45.58	46.65	36.67	38.62	44.05	44.06
Transformer with copy	59.56	57.91	39.93	45.10	59.27	52.25
Pretrained BART	**79.32**	**78.53**	62.30	61.82	**77.49**	**78.03**
Pipeline	65.56	70.03	**67.64**	**62.45**	67.27	74.16

Table 7.2: Automatic evaluation results for the task of idiomatic sentence generation (**s2i**) and idiomatic sentence paraphrasing (**i2s**).

appropriate. For GRUEN metric aiming to measure text quality, its correlation scores with fluency and overall meaning are quite low. Therefore, more reliable automatic evaluation methods are needed.

Error analysis. For task of idiomatic sentence generation, the primary challenge is in identifying the appropriate IE, which is the hardest when the IE is highly non-compositional (e.g., *bird of passage* in Table 7.6). The examples are presented in Table 7.6. For the task of idiomatic sentence paraphrasing, one challenge is the difficulty of choosing the correct sense of the idiom. As is shown in Table 7.7, all the baseline models were unable to generate the correct literal phrases for "alpha and omega", which have two senses: the beginning and the end; the principal element. Also, we noticed that strong baseline models of pretrained BART and the pipeline model tend to use a short but inaccurate literal phrase when the correct one

Corr	Context		Target		Fluency		Overall	
	s2i	i2s	s2i	i2s	s2i	i2s	s2i	i2s
BLEU	0.27	0.17	0.56	0.28	0.09	0.02	0.64	0.29
SARI	0.21	0.17	0.61	0.40	-0.02	-0.01	0.61	0.39
GRUEN	-0.18	-0.07	-0.11	0.12	0.23	0.15	-0.18	0.11

Table 7.3: Instance-level Spearman's correlations between human and automatic evaluation for pretrained BART.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	SARI	GRUEN	Perplexity
Seq2Seq	25.16	48.26	22.90	47.21	41.46	24.13	32.25	4.24
Seq2Seq with copy	38.02	66.11	40.37	74.04	68.21	43.02	27.79	24.43
Transformer	45.58	60.22	42.82	60.59	68.68	36.67	44.05	4.00
Transformer with copy	59.56	68.34	55.72	69.38	79.53	39.93	59.27	4.12
Pretrained BART	**79.32**	**83.95**	**77.16**	**84.20**	**83.41**	62.30	**77.49**	3.88
Pipeline	65.56	74.44	62.96	74.56	78.02	**67.64**	67.27	**3.4**

Table 7.4: Performance comparison of baselines for idiomatic sentence generation

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	SARI	GRUEN	Perplexity
Seq2Seq	42.96	62.43	40.46	62.54	59.36	33.89	33.45	9.54
Seq2Seq with copy	47.58	71.67	50.20	76.77	77.23	49.69	32.84	21.85
Transformer	46.65	60.90	43.34	61.39	69.82	38.62	44.06	10.59
Transformer with copy	57.91	68.44	54.97	69.59	79.17	45.10	52.25	4.61
Pretrained BART	**78.53**	**84.64**	**77.21**	**84.95**	**85.36**	61.82	**78.03**	5.35
Pipeline	70.03	78.50	68.39	78.90	83.65	**62.45**	74.16	**4.25**

Table 7.5: Performance comparison of baselines for idiomatic sentence paraphrasing

is long. Paraphrasing of "the bird of passage" in Table 7.7 is an example.

Applications: Research in the proposed tasks has many potential practical applications. 1) An idiomatic sentence paraphrasing tool would be of importance in several language processing settings encountered by by humans and machines. The non-literal and stylized meaning of multi-word expressions (MWE) in general and idioms in particular, pose two broad kinds of challenges. First, they affect readability in target populations. For instance, despite their intact structural language competence, individuals with Asperger syndrome and more broadly those with autism spectrum disorder are known to experience significant challenges understanding figurative language (idioms) in their native language [31]. It is also widely acknowledged that idiomatic expressions are some of the hardest aspects of language acquisition and processing for second language learners [32, 33, 34]. Moreover, natural language processing systems are known to be negatively impacted by idioms in text ([13, 14, 15] shown the negative impact of idioms and metaphors on machine translation leading to awkward or incorrect translations from English to other languages). Fruitful results of this task can lead to a system capable of recognizing and interpreting IEs in unrestricted text in a central component of any real-world NLP application (e.g., information retrieval, machine translation, question answering, information extraction, and opinion mining).2) A realistic application of the idiomatic sentence generation task would be for computer-aided style checking, where a post-processing tool could suggest a list of idioms to replace a literal phrase in a sentence. 3) True integration with an external NLP application would require combining the first step of IE identification followed by paraphrasing as done in [15], which will require a combination of the paraphrasing with identification, and can be a future direction for research.

Attribute	high non-compositionality
Literal sentence	Joe , being one who is here today and gone tomorrow , stayed the night , had some rest and ate some food and left early the next morning .
Reference	Joe , being the bird of passage he is , stayed the night , had some rest and ate some food and left early the next morning .
Seq2Seq	First , being one , and putting the project going to be joined the ones , had some ice row and creating some people and creating some expensive of both the time .
Transformer	joe , being one who is here today and gone tomorrow , kept the night , had some rest and punched some food a great early .
Seq2Seq with copy	joe , being the bird of he he , , , , , , , some some some some and and and and the .
Transformer with copy	joe , being one who is here today and gone tomorrow , stayed the night , had a rest and ate food left the next early .
Pretrained BART	Joe, being one who is here today and gone tomorrow, stayed the night, had some rest and ate some food and left early the next morning.
Pipeline	cool heels joe, being one who is here today and gone tomorrow, stayed the night, and ate some food and left early the next morning.
Attribute	**multiple meaning**
Literal sentence	My life starts from you and ends at you , so you are my first and my last .
Reference	My life starts from you and ends at you , so you are my alpha and omega .
Seq2Seq	My friend from you and offensive , and yet you are my dream and my loved .
Transformer	my life starts from you and anything at you , so you are my first sight and my last .
Seq2Seq with copy	my life starts from you and at you you you you you you my my and .
Transformer with copy	My life starts from you and ends at you , so you are my first and my last .
Pretrained BART	My life starts from you and ends at you , so you are my first and my last.
Pipeline	Close the books, so you are my my first and my last.
Attribute	**high non-compositionality**
Literal sentence	You can't delay making a decision any longer , you need to make up your mind .
Reference	You can't sit on the fence any longer , you need to make up your mind .
Seq2Seq	You can't be in the obsession any night , you need to make up your plans .
Transformer	you can't delay making a decision of any longer , you need to make your mind your mind .
Seq2Seq with copy	you can't sit sit the fence any , , you need to to up your .
Transformer with copy	you can't delay making a decision any longer , you need to make up your mind .
Pretrained BART	You can't delay making a decision any longer, you need to make up your own mind.
Pipeline	You can't delay making a decisione any longer, you make your mind.
Attribute	**low non-compositionality**
Literal sentence	Finding the ruins of Babylon was the archaeologist 's greatest find .
Reference	Finding the ruins of Babylon was the archaeologist 's treasure trove .
Seq2Seq	Missing the aftermath of pouring down the cake 's share of the city .
Transformer	catching up with silver lining of the challenges 's volatility .
Seq2Seq with copy	finding the ruins of unk was the 's 's trove .
Transformer with copy	finding the ruins of babylon was the archaeologist 's greatest silver spoons .
Pretrained BART	Finding the ruins of Babylon was the archaeologist's greatest find.
Pipeline	Finding the ruins of babylon was the archaeologist' treasure trove.

Table 7.6: Samples of generated idiomatic sentences. Text in blue represents the idiomatic expressions correctly included in the outputs; text in red represents the literal counterparts in the input sentences. text in green represents the idioms that are poorly generated.

Attribute	high non-compositionality
Idiomatic sentence	Joe , being the bird of passage he is , stayed the night , had some rest and ate some food and left early the next morning .
Reference	Joe , being one who is here today and gone tomorrow , stayed the night , had some rest and ate some food and left early the next morning .
Seq2Seq	And , sitting the part of the Bieber he is , seemed the morning , he some smart and wound problems so well and gives early at the next morning .
Transformer	joe , being the guards of nowhere he is , the night the night , and had some dealers and left the morning left the next morning .
Seq2Seq with copy	joe , being one who here today and tomorrow tomorrow stayed stayed night , had some and and and and and left next next next .
Transformer with copy	joe , being the bird of energy is stayed , stayed the night , some rest and ate ate some food left the next morning .
Pretrained BART	Joe, being the traveler he is, stayed the night, had some rest and ate some food and left early the next morning.
Pipeline	joe, being the person he is, stayed the night, had some rest and ate some food and left early the next morning.
Attribute	**multiple meaning**
Idiomatic sentence	My life starts from you and ends with you , so you are my alpha and omega .
Reference	My life starts from you and ends with you , so you are my first and my last .
Seq2Seq	My life dreams from you and read your family at you , so you are .
Transformer	my life starts from you and learn at you , so you are my most important part .
Seq2Seq with copy	my life starts from you ends ends you , so you my my my my last last last .
Transformer with copy	my life starts from you and ends with you , so you are my wish and omega .
Pretrained BART	My life starts from you and ends with you, so you are my most important part.
Pipeline	My life starts from you and ends with you, so you are my most important part.
Attribute	**high non-compositionality**
Idiomatic sentence	You can't sit on the fence any longer , you need to make up your mind .
Reference	You can't delay making a decision any longer , you need to make up your mind .
Seq2Seq	You can't wait on the money any rival , you need to make up your energy .
Transformer	you can't sit on the ? any longer , you need to make up your mind .
Seq2Seq with copy	you can't delay making any any any , you need to make your your mind .
Transformer with copy	you ca n't sit on the troublesome any longer , you need to make your mind .
Pretrained BART	You can't be indecisive any longer, you need to make up your mind.
Pipeline	You can't stay on the fence any longer, you need to make up your mind.
Attribute	**low non-compositionality**
Idiomatic sentence	Finding the ruins of Babylon was the archaeologist 's treasure trove .
Reference	Finding the ruins of Babylon was the archaeologist 's greatest find .
Seq2Seq	Edward the trap of nature was the racial out of Robert .
Transformer	finding and hide of confiement was shocking 's legal code .
Seq2Seq with copy	finding the ruins of unk was the unk 's greatest find .
Transformer with copy	finding the ruins of babylon was the archaeologist's family members .
Pretrained BART	Finding the ruins of Babylon was the archaeologist's greatest find.
Pipeline	Finding the ruins of babylon was the archaeologist's trove.

Table 7.7: Samples of generated literal sentences. Text in red represents the appropriate literal phrases included in the outputs. Text in blue represents the idioms in the input sentences. Text in green represents the literal phrases that are poorly generated.

CHAPTER 8: ENLARGED CORPUS

Based on the previous experiment results on the parallel dataset, it is shown that the pretrained BART model has a quite strong ability of generating literal counterparts for idiomatic sentences. Therefore, considering that the IEs and parallel examples collected from the online resources are on a small scale, we use BART to enlarge the number of parallel instances using the publicly available MAGPIE dataset [30], a collection of sentences with IEs collected from the British National Corpus. With the help of the enlarged corpus, we could also observe an improvement of performance on our task of idiomatic sentence generation for baseline models.

8.1 MAGPIE CORPUS

MAGPIE corpus is the largest corpus of sense-annotated IEs to date using a crowdsourced annotation approach. Sentences in MAGPIE corpus are extracted from the British National Corpus. Therefore, examples in this corpus are all real examples in the wild, which is an advantage over the synthetic examples.

To leverage the examples in MAGPIE corpus, we first excluded the sentences with IEs used in a literal sense using the labels provided in the dataset. Then we excluded sentences longer than 30 words to keep them comparable to the manually created dataset and to avoid any long-range dependency challenges for generation. This resulted in 1536 idioms and 17000 idiomatic sentences over 1536 IEs (average of 11.07 sentences per IE).

8.2 HUMAN-IN-THE-LOOP SELF-TRAINING

Because of the large number of idiomatic sentences in the MAGPIE dataset, we generate the corresponding literal examples following a human-in-the-loop self-training process using a large pretrained BART model (BART). We first use the parallel idiomatic-literal sentences from previous parallel dataset as the training corpus for BART, treating the idiomatic sentences as input to the model that is fine-tuned to generate the corresponding literal sentences. After fine-tuning, BART is then used to generate literal counterparts for idiomatic examples in the MAGPIE dataset. Due to the significant overlap between the input and the output sentences in the training set, a vast majority of the BART output for the idiomatic sentences in the MAGPIE dataset was identical to the input. Therefore, we extract the input-output pairs that are different and manually 'correct' them following the same process

Statistics	# of instances	Avg. # of words
Idioms	2078	3.4
Definitions	2117	10.4
Idiomatic sent	22170	18.6
Literal sent	22170	18.1

Table 8.1: Statistics of our parallel corpus.

Idiom	**up and running**
Idiomatic	The Legal Centre was up and running and ready for business.
Literal	The Legal Centre was working again and ready for business.
Idiom	**new blood**
Idiomatic	My biggest regret is that there 'll be no new blood keeping the spirit going.
Literal	My biggest regret is that there 'll be no new reforming members keeping the spirit going.
Idiom	**foot the bill**
Idiomatic	In many cases, genuine customers footed the bill under old laws governing cheques.
Literal	In many cases, genuine customers covered the cost under old laws governing cheques.
Idiom	**under fire**
Idiomatic	This was the first time a Social Work Department in Scotland had come under such public fire for the same reasons.
Literal	This was the first time a Social Work Department in Scotland had come under such public harsh criticism for the same reasons.

Table 8.2: Samples of generated literal sentences. Text in red represents the literal phrases included in the outputs. Text in blue represents the idioms in the input sentences.

for creating the previous parallel dataset. These manually created idiomatic-literal pairs are then added into the training set and used for further fine-tuning the BART model. By using this human-in-the-loop method for two rounds, we successfully generated 17000 idiomatic-literal sentence pairs from MAGPIE dataset.

8.3 RESULTS

After enlargement, 17000 more sentence pairs are added in the the parallel corpus. We summarize the corpus statistics in Table 8.1. Some examples of the generated sentence pairs are provided in Table 8.2. From the samples in Table 8.2, we could learn that literal sentences generated by the pretrained BART model are of high quality, which also guarantees the quality of enlarged parallel dataset.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Perplexity
Seq2Seq	55.51	70.69	54.42	71.07	67.48	7.69
Transformer	60.29	63.15	53.35	64.55	70.55	7.71
Seq2Seq with copy	55.86	70.81	56.72	73.70	78.75	7.24
Transformer with copy	61.87	68.80	57.44	69.83	78.21	7.27

Table 8.3: The improved performance of baselines and our model based on the enlarged corpus.

Similar to the previous parallel dataset, we also annotated the idioms with BIO labels [86] to mark the ground truth span of idioms in the idiomatic sentences and also annotated the literal phrases to mark the ground truth span of corresponding literal counterparts. The annotation is completed automatically by detecting the different parts between the idiomatic sentences and corresponding literal counterparts.

8.4 IMPROVEMENT

With the help of the enlarged corpus, we obtained more samples to train different models. Results of experiments for idiomatic sentence generation are presented in Table 8.3. From the results shown in Table 8.3, we can see the improvement performance with the help of the enlarged corpus. All the baseline models have a better performance compared with the results shown in Table 7.4.

CHAPTER 9: CONCLUSIONS AND FUTURE WORKS

9.1 CONCLUSION

To conclude, in this study, we proposed two new tasks: idiomatic sentence generation and paraphrasing. We also presented PIE, the first parallel idiom corpus. We also propose a novel approach for these tasks, which retrieves the appropriate idiom for a given literal sentence, extracts the span of the sentence to be replaced by the idiom, and generates the idiomatic sentence by using a large pretrained language model to combine the retrieved idiom and the remainder of the sentence. For idiomatic sentence paraphrasing, the definition of the idiom in the given idiomatic sentence is first retrieved. Then the idiom in the sentence is extracted and finally the literal counterpart is generated by a large pretrained language model.

We benchmark existing end-to-end trained neural network models and a pipeline method on PIE and analyze their performance for our tasks. Our experiments and analyses reveal the competence and shortcomings of available methods, underscoring the need for continued research on processing idiomatic expressions.

Inspired by the good performance of the pretrained BART model, we also utilized it to enlarge our parallel dataset with the help of MAGPIE corpus. Finally, a larger parallel dataset with 2078 idioms and 22170 idiomatic-literal sentence pairs is created. This enlarged dataset is also shown to be useful for improving performance for idiomatic sentence generation.

9.2 LIMITATIONS

Model. Based on the previous evaluation, we could know that due to high non-compositionality of idioms it is very difficult to retrieve the appropriate idioms for the task of idiomatic sentence generation, which will further influence the quality of finally generated idiomatic sentences. Besides, we can also observe that the pretrained BART model just copied the input into the output sometimes due to the high similarity between input and output in the training set, which makes the fine-tuned BART a simple copy model.

Evaluation. The correlation presented in Table 7.3 showed that current automatic evaluation metrics are not good enough. Synonymous idioms and literal phrases cannot be measured by current evaluation metrics such as BLEU and SARI. In addition, the high similarity between input and output also made the scores of automatic evaluation metrics too high and thus mitigated the difference between different models' performance.

9.3 FUTURE WORK

Based on the limitations discussed above, there are many possibilities for improving performance through more extensive exploration of richer model architectures and using more reliable evaluation methods, especially considering that currently evaluation metrics cannot evaluate the results and performance perfectly because of the high overlapping between input and output.

The second direction of future work is to apply our work in adversarial example generation. Idioms are naturally of high non-compositionality and thus are difficult for language model to understand only based on word embeddings. Therefore, texts including idioms are expected to confuse current neural network models which are build for many text classification tasks, for example sentiment analysis, natural language inference, paraphrase identification and etc. Current methods of adversarial example generation rely on the model to be attacked more or less. Some of them rely on the model's output [87, 88, 89] and others rely on the model's structure and gradient information [90, 91, 92]. However, our work is able to be further developed into a model-agnostic method because the idioms inserted are potentially difficult to process for all the models.

Some works have already pointed out the influence idioms have on current neural models. For example, [93] created a large-scale dataset for idiom translation. The machine translation results showed that currently neural machine translation models have an obviously poorer performance on idiom translation compared with literal sentence translation. Therefore, our method for idiomatic sentence paraphrasing could be used to first transfer the idioms into their literal counterparts. Then, instead of directly translating idiomatic sentences, the NMT models could translate the corresponding literal sentences for a better performance.

Another direction our work can be applied into is paraphrase and style transfer. For these tasks, our work could be used as a method of data augmentation to generate more training examples for these tasks. Considering that our work on idiomatic sentence generation only inserts idioms into the original sentences and retain the original semantic meanings, idiomatic sentences generated by our method are essentially paraphrases for original sentences. Therefore, our work could also be used for data augmentation for text generation tasks like paraphrase generation and style transfer.

REFERENCES

[1] G. Nunberg, I. A. Sag, and T. Wasow, "Idioms," *Language*, vol. 70, no. 3, pp. 491–538, 1994.

[2] A. Wray and M. R. Perkins, "The functions of formulaic language: An integrated model," *Language & Communication*, vol. 20, no. 1, pp. 1–28, 2000.

[3] S. A. Sprenger, "Fixed expressions and the production of idioms," Ph.D. dissertation, Radboud University Nijmegen Nijmegen, 2003.

[4] A. Pawley and F. H. Syder, "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency," in *Language and communication*. Routledge, 2014, pp. 203–239.

[5] N. Schmitt and D. Schmitt, *Vocabulary in language teaching*. Cambridge university press, 2020.

[6] R. Simpson and D. Mendis, "A corpus-based study of idioms in academic speech," *Tesol Quarterly*, vol. 37, no. 3, pp. 419–441, 2003.

[7] R. Jackendoff, "The boundaries of the lexicon," *Idioms: Structural and psychological perspectives*, pp. 133–165, 1995.

[8] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, "Longman grammar of written and spoken english," *Harlow: Longman*, 1999.

[9] D. Kerbel and P. Grunwell, "Idioms in the classroom: An investigation of language unit and mainstream teachers' use of idioms," *Child Language Teaching and Therapy*, vol. 13, no. 2, pp. 113–123, 1997.

[10] R. Baldwin, M. Cave, and M. Lodge, *The Oxford handbook of regulation*. Oxford university press, 2010.

[11] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword expressions: A pain in the neck for nlp," in *International conference on intelligent text processing and computational linguistics*. Springer, 2002, pp. 1–15.

[12] P. Rayson, S. Piao, S. Sharoff, S. Evert, and B. V. Moirón, "Multiword expressions: hard going or plain sailing?" *Language Resources and Evaluation*, vol. 44, no. 1-2, pp. 1–5, 2010.

[13] G. Salton, R. Ross, and J. Kelleher, "An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese," 2014.

[14] Y. Shao, R. Sennrich, B. Webber, and F. Fancellu, "Evaluating machine translation performance on chinese idioms with a blacklist method."

[15] E. Shutova, S. Teufel, and A. Korhonen, "Statistical metaphor processing," *Computational Linguistics*, vol. 39, no. 2, pp. 301–353, 2013.

[16] B. B. Klebanov, K. Knight, and D. Marcu, "Text simplification for information-seeking applications," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2004, pp. 735–747.

[17] M. Constant, G. Eryiğit, J. Monti, L. Van Der Plas, C. Ramisch, M. Rosner, and A. Todirascu, "Multiword expression processing: A survey," *Computational Linguistics*, vol. 43, no. 4, pp. 837–892, 2017.

[18] M. Baker, *In other words: A coursebook on translation*. Routledge, 2018.

[19] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence to sequence models," in *Proceedings of the Workshop on Stylistic Variation*, 2017, pp. 10–19.

[20] H. Gong, S. Bhat, L. Wu, J. Xiong, and W.-m. Hwu, "Reinforcement learning based text style transfer without parallel training corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3168–3180.

[21] E. Shutova, "Models of metaphor in nlp," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 688–697.

[22] A. Savary, C. Ramisch, S. R. Cordeiro, F. Sangati, V. Vincze, B. Qasemi Zadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova et al., "The parseme shared task on automatic identification of verbal multiword expressions," in *The 13th Workshop on Multiword Expression at EACL*, 2017, pp. 31–47.

[23] C. Liu and R. Hwa, "A generalized idiom usage recognition model based on semantic compatibility," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6738–6745.

[24] S. Muresan, R. Gonzalez-Ibanez, D. Ghosh, and N. Wacholder, "Identification of nonliteral language in social media: A case study on sarcasm," *Journal of the Association for Information Science and Technology*, vol. 67, no. 11, pp. 2725–2737, 2016.

[25] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 5, pp. 1–22, 2017.

[26] D. Ghosh, A. R. Fabbri, and S. Muresan, "Sarcasm analysis using conversation context," *Computational Linguistics*, vol. 44, no. 4, pp. 755–792, 2018.

[27] E. Shutova, "Automatic metaphor interpretation as a paraphrasing task," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 1029–1037.

[28] P. Cook, A. Fazly, and S. Stevenson, "The vnc-tokens dataset," in *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, 2008, pp. 19–22.

[29] I. Korkontzelos, T. Zesch, F. M. Zanzotto, and C. Biemann, "Semeval-2013 task 5: Evaluating phrasal semantics," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 39–47.

[30] H. Haagsma, J. Bos, and M. Nissim, "Magpie: A large corpus of potentially idiomatic expressions," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 279–287.

[31] T. Kalandadze, C. Norbury, T. Nærland, and K.-A. B. Næss, "Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review," *Autism*, vol. 22, no. 2, pp. 99–117, 2018.

[32] J. Liontas, "Context and idiom understanding in second languages," *EUROSLA yearbook*, vol. 2, no. 1, pp. 155–185, 2002.

[33] N. C. Ellis, R. Simpson-Vlach, and C. Maynard, "Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and tesol," *Tesol Quarterly*, vol. 42, no. 3, pp. 375–396, 2008.

[34] E. Canut, J. Delahaie, and M. Husianycia, "Vous avez dit falc? pour une adaptation linguistique des textes destinés aux migrants nouvellement arrivés," *Langage et societe*, no. 3, pp. 171–201, 2020.

[35] G. Freyhoff, G. Hess, L. Kerr, E. Menzel, B. Tronbacke, and K. Van Der Veken, "Make it simple, european guidelines for the production of easy-to-read information for people with learning disability for authors, editors, information providers, translators and other interested persons," *International League of Societies for Persons with Mental Handicap European Association, Brussels*, 1998.

[36] K. A. Schriver, "Plain language in the us gains momentum: 1940–2015," *Ieee transactions on professional communication*, vol. 60, no. 4, pp. 343–383, 2017.

[37] S. O'Brien, "Controlling controlled english. an analysis of several controlled language rule sets," *Proceedings of EAMT-CLAW*, vol. 3, no. 105-114, p. 33, 2003.

[38] W. Xu, C. Callison-Burch, and C. Napoles, "Problems in current text simplification research: New data can help," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 283–297, 2015.

[39] S. Agrawal and M. Carpuat, "Multitask models for controlling the complexity of neural machine translation," in *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, 2020, pp. 136–139.

[40] A. Fazly, P. Cook, and S. Stevenson, "Unsupervised type and token identification of idiomatic expressions," *Computational Linguistics*, vol. 35, no. 1, pp. 61–103, 2009.

[41] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "Ppdb: The paraphrase database," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 758–764.

[42] W. Lan, S. Qiu, H. He, and W. Xu, "A continuously growing dataset of sentential paraphrases," *arXiv preprint arXiv:1708.00391*, 2017.

[43] J. Wieting and K. Gimpel, "Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations," *arXiv preprint arXiv:1711.05732*, 2017.

[44] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, "Neural paraphrase generation with stacked residual lstm networks," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2923–2934.

[45] A. Gupta, A. Agarwal, P. Singh, and P. Rai, "A deep generative framework for paraphrase generation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[46] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1875–1885.

[47] Q. Yang, D. Shen, Y. Cheng, W. Wang, G. Wang, L. Carin et al., "An end-to-end generative architecture for paraphrase generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3123–3133.

[48] Y. Gu, Z. Wei et al., "Extract, transform and filling: A pipeline model for question paraphrasing based on template," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 109–114.

[49] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Advances in neural information processing systems*, 2017, pp. 6830–6841.

[50] S. Rao and J. Tetreault, "Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer," *arXiv preprint arXiv:1803.06535*, 2018.

[51] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 897–908.

[52] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.

[53] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1587–1596.

[54] K.-H. Zeng, M. Shoeybi, and M.-Y. Liu, "Style example-guided text generation using generative adversarial transformers," *arXiv preprint arXiv:2003.00674*, 2020.

[55] J. Xu, X. Sun, Q. Zeng, X. Zhang, X. Ren, H. Wang, and W. Li, "Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 979–988.

[56] J. Li, R. Jia, H. He, and P. Liang, "Delete, retrieve, generate: a simple approach to sentiment and style transfer," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1865–1874.

[57] A. Sudhakar, B. Upadhyay, and A. Maheswaran, ""transforming" delete, retrieve, generate approach for controlled text style transfer," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3260–3270.

[58] K. Abe, K. Sakamoto, and M. Nakagawa, "A computational model of the metaphor generation process," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 28, no. 28, 2006.

[59] S. Mohammad, E. Shutova, and P. Turney, "Metaphor as a medium for emotion: An empirical study," in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 2016, pp. 23–33.

[60] A. Terai and M. Nakagawa, "A computational system of metaphor generation with evaluation mechanism," in *International Conference on Artificial Neural Networks*. Springer, 2010, pp. 142–147.

[61] E. Ovchinnikova, V. Zaytsev, S. Wertheim, and R. Israel, "Generating conceptual metaphors from proposition stores," *arXiv preprint arXiv:1409.7619*, 2014.

[62] R. Mao, C. Lin, and F. Guerin, "Word embedding and wordnet based metaphor identification and interpretation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1222–1231.

[63] Z. Yu and X. Wan, "How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 861–871.

[64] K. Stowe, L. Ribeiro, and I. Gurevych, "Metaphoric paraphrase generation," *arXiv preprint arXiv:2002.12854*, 2020.

[65] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 1353–1361.

[66] X. Zhang and M. Lapata, "Sentence simplification with deep reinforcement learning," *arXiv preprint arXiv:1703.10931*, 2017.

[67] S. Wubben, E. Krahmer, and A. van den Bosch, "Sentence simplification by monolingual machine translation," 2012.

[68] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu, "Exploring neural text simplification models," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 85–91.

[69] S. Zhao, R. Meng, D. He, S. Andi, and P. Bambang, "Integrating transformer and paraphrase rules for sentence simplification," *arXiv preprint arXiv:1810.11193*, 2018.

[70] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 86–96.

[71] J. Ive, N. Viani, J. Kam, L. Yin, S. Verma, S. Puntis, R. N. Cardinal, A. Roberts, R. Stewart, and S. Velupillai, "Generation and evaluation of artificial mental health records for natural language processing," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[72] R. W. Gibbs Jr, "Idioms and formulaic language," in *The Oxford handbook of cognitive linguistics*, 2007.

[73] S. A. Sprenger, W. J. Levelt, and G. Kempen, "Lexical access during the production of idiomatic phrases," *Journal of memory and language*, vol. 54, no. 2, pp. 161–184, 2006.

[74] S. Sprenger, A. la Roi, and J. Van Rij, "The development of idiom knowledge across the lifespan," *Frontiers in Communication*, vol. 4, p. 29, 2019.

[75] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[76] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, no. Mar, pp. 693–723, 2007.

[77] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[79] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," *arXiv preprint arXiv:1808.10792*, 2018.

[80] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[81] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004. [Online]. Available: https://www.aclweb.org/anthology/W04-1013 pp. 74–81.

[82] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[83] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the second workshop on statistical machine translation*, 2007, pp. 228–231.

[84] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.

[85] W. Zhu and S. Bhat, "Gruen for evaluating linguistic quality of generated text," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 94–108.

[86] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176.

[87] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? a strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8018–8025.

[88] D. Li, Y. Zhang, H. Peng, L. Chen, C. Brockett, M.-T. Sun, and B. Dolan, "Contextualized perturbation for textual adversarial attack," *arXiv preprint arXiv:2009.07502*, 2020.

[89] R. Maheshwary, S. Maheshwary, and V. Pudi, "Generating natural language attacks in a hard label black box setting," *arXiv preprint arXiv:2012.14956*, 2020.

[90] L. Song, X. Yu, H.-T. Peng, and K. Narasimhan, "Universal adversarial attacks with natural triggers for text classification," *arXiv preprint arXiv:2005.00174*, 2020.

[91] B. Wang, H. Pei, B. Pan, Q. Chen, S. Wang, and B. Li, "T3: Tree-autoencoder constrained adversarial text generation for targeted attack," *arXiv preprint arXiv:1912.10375*, 2019.

[92] Y. Cheng, L. Jiang, and W. Macherey, "Robust neural machine translation with doubly adversarial inputs," *arXiv preprint arXiv:1906.02443*, 2019.

[93] M. Fadaee, A. Bisazza, and C. Monz, "Examining the tip of the iceberg: A data set for idiom translation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.